

# DeepDive 科普报道

## 简介

DeepDive 是一个可以通过机器学习技术利用特定领域知识，并且通过用户反馈提高分析质量的开源知识提取系统。

其目标为帮助用户从数据中提取实体和关系，并对事实进行推理。

DeepDive 可以处理结构化/非结构化、干净/带噪声的数据，并将结果输出至数据库。

主要面向从互联网非结构化数据中抽取结构化信息，做一系列后处理，构建知识库并抽取关系等。

DeepDive 的特点为：

1. 意识到数据存在噪音（不精确性），通过为每个断言设置置信度予以校正
2. 数据源多样性，如文档、网页、图表等
3. 可以通过使用已有的领域知识指导推理，接受用户反馈，提高预测的质量
4. 使用 Distant supervision 技术，只需少量/甚至不需要训练数据
5. 是一个可拓展的高性能推理和学习引擎

## 要求

用户需要：

1. 对 SQL 和 Python 熟悉，以便使用 Deepdive 或集成 Deepdive 的工具开发
2. 需要对 DeepDive 进行改善的用户还需了解知识库构建（Knowledge base construction）、关系提取（Relation extraction）、Distant supervision、概率推理（Probabilistic inference）、因子图（factor graphs）等基础背景知识

# 数据模型

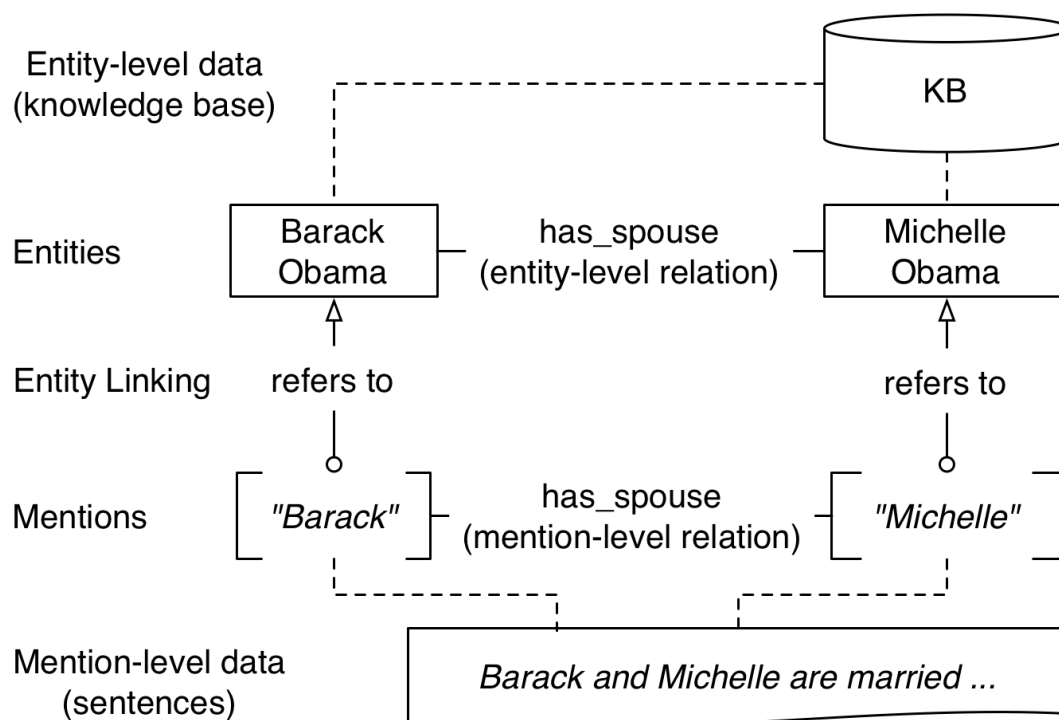


图 1 数据模型

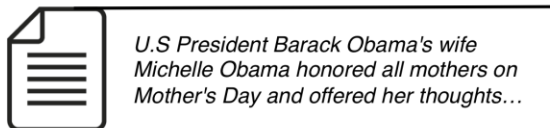
- 实体：现实中存在的事物，如奥巴马
- Mention：对实体的一个引用，如“奥巴马”三个字
- 实体级关系：实体间的关系
- Mention 级关系：Mention 间的关系
- 实体级数据：实体级关系的集合，如 Freebase（知识库）中的关系
- Mention 级数据：包含 Mention 的数据，如"Barack and Michelle are married"这个句子
- 实体耦合：Mention 与实体的映射

## 工作流程

这里以“提取文本中 Mention 级关系”为例，介绍 DeepDive 的工作流程。

0. 数据预处理。将数据（如文本）载入数据库并解析，获得语句级的信息，包括 POS tags, named entity tags 等。
1. 特征提取。通过开发者编写的提取器，将数据数据转化为关系表示，称为 **evidence**
2. 建立因子图。开发者类 SQL 语言，声明推理规则。
3. 学习与统计推理。DeepDive 自动进行，可以计算出某特定断言的可能性。

图 2 为相应流程的示意图



## 0. Data Preprocessing

Sentence	POS	NER
[U.S,President,Barack,Obama,'s,wife,Michelle,Obama...]	[NNP,NNP,NNP,NNP,POS,NN,NNP,NNP...]	[LOC,O,PERSON,PERSON,O,O,PERSON,PERSON...]

## 1. Feature Extraction

Mentions		CandidateRelations		
Mention	Type	Subject	Object	has_spouse
U.S	LOC	Barack Obama	Michelle Obama	
Barack Obama	PERSON			
Michelle Obama	PERSON			

Features		
Subject	Object	feature
Barack Obama	Michelle Obama	's wife

## 2. Writing Inference Rule

```
inference_rule_1:{
  input_query: ""
  SELECT CandidateRelations.has_spouse, Features.feature
  FROM   CandidateRelations, Features
  WHERE  CandidateRelations.Subject=Features.Subject
        AND  CandidateRelations.Object=Features.Object""
  function: "IsTrue(CandidateRelations.has_spouse)"
  weight: "(Features.feature)"
}
```

图 2 流程示意图

# 开发者

主持人: Christopher Ré (Stanford University)

团队成员:

- Zifei Shan
- Feiran Wang
- Sen Wu
- Jaeo Shin
- Amir Abbas Sadeghian
- Ce Zhang
- Matteo Riondato.

# 链接

- <http://deepdive.stanford.edu/index.html>
- [http://cs.stanford.edu/people/chrismre/papers/deepdive\\_vlds.pdf](http://cs.stanford.edu/people/chrismre/papers/deepdive_vlds.pdf)